Lessons Learned The Hard Way

Bob Lucas

USC Information Sciences Institute

June 26, 2002

This is all Fred's fault!

He made me give this talk ☺

What Howard Frank's Wife Said

- You should have three points
- Any less, and you have nothing to say
- Any more, and you have nothing important

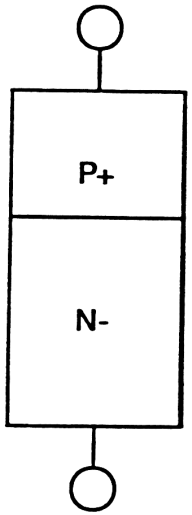
What I'm Going To Talk About

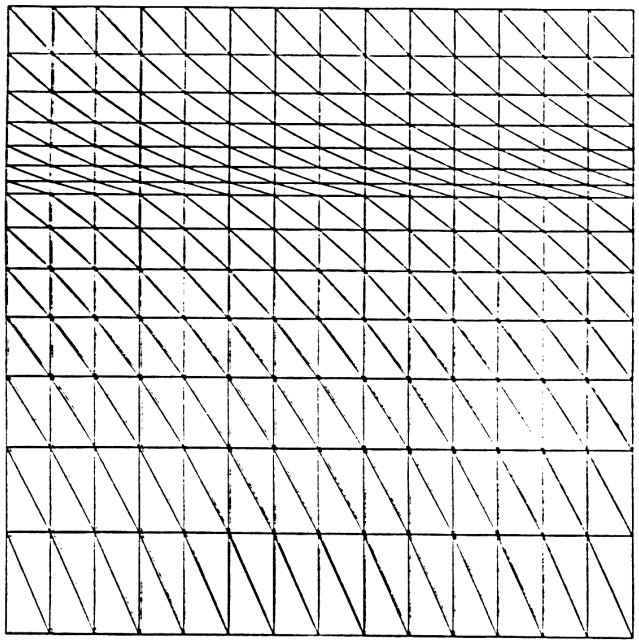
- Parallel applications
- Programming models
- The "Vision Thing"

Application War Story #1

- My main thesis project: Parallel PISCES
- Baseline is PISCES 2B
  - 2D transistor device modeling code
  - Bottleneck is sparse matrix solver
  - Platform of choice is the all-powerful VAX 11/780
- Research Question:
  - See if parallel processing addresses this major computational bottleneck in electrical engineering.

Trivial PISCES Example





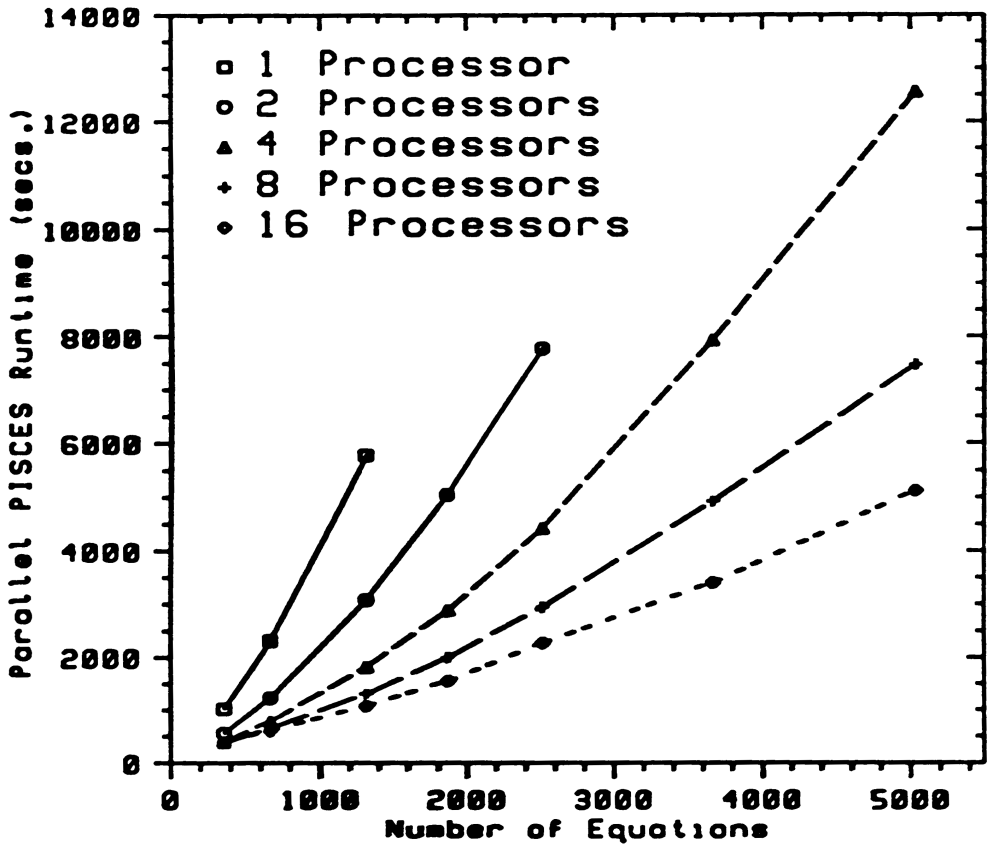
Simulation grid for 15 by 15 diode

PISCES Input Deck


```
title square pn diode
mesh   rect nx=15 ny=15
x.mesh location=0.0 node=1  ratio=1
x.mesh location=1.0 node=15 ratio=1
y.mesh location=0.0 node=1  ratio=1
y.mesh location=0.3 node=8  ratio=0.8
y.mesh location=1.0 node=15 ratio=1.2
region num=1 silicon ix.lo=1 ix.hi=15 iy.lo=1 iy.hi=15
elec num=1 ix.lo=1 ix.hi=15 iy.lo=1 iy.hi=1
elec num=2 ix.lo=1 ix.hi=15 iy.lo=15 iy.hi=15
doping reg=1 n.type conc=1e15 uniform
doping reg=1 p.type conc=1e19 gauss
+ x.l=0 x.r=1 y.top=0 y.bot=0 junc=0.3
symb   newton  cube  carr=2
method rhsnorm xnorm autonr
models temp=300 srh auger conmob fldmob
solve   init  vstep=0.1 nsteps=3 elect=1
end
```

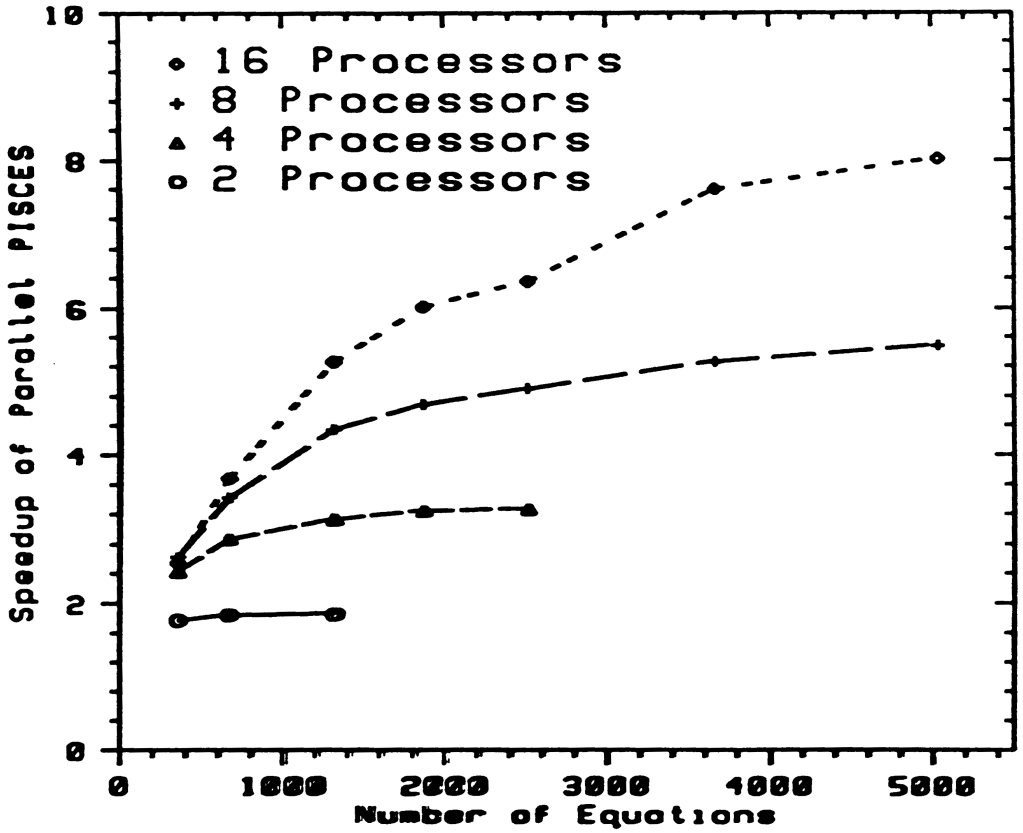
PISCES command file for 15 X 15 diode

Parallel PISCES Run Time



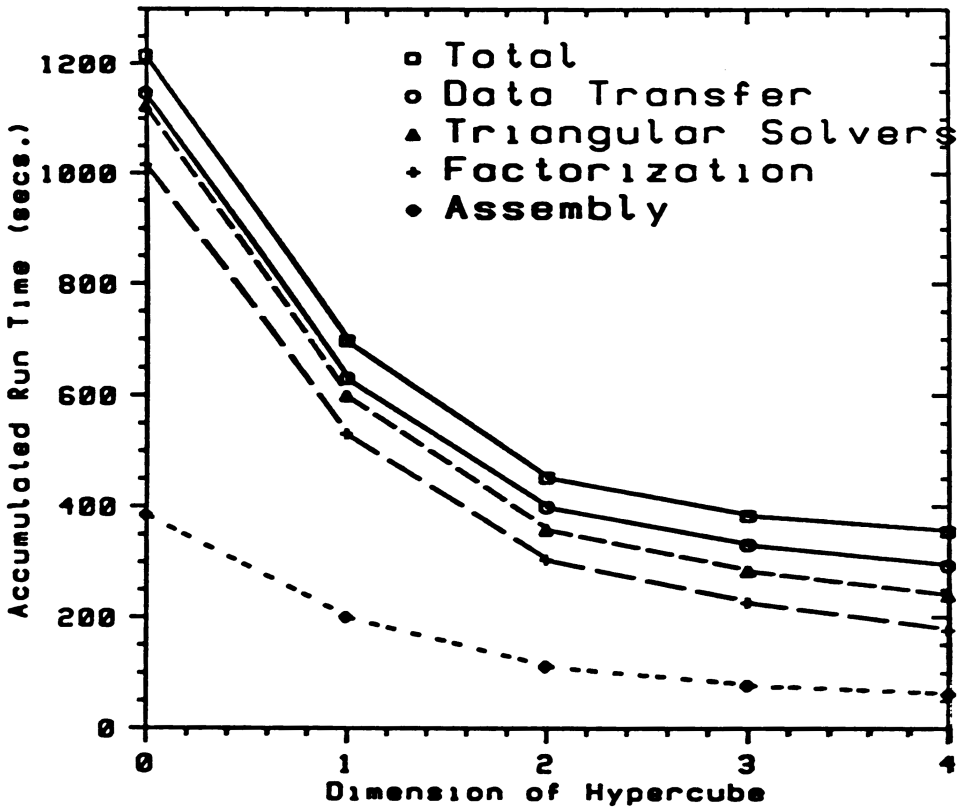
Parallel PISCES Run Time

Parallel PISCES Speedup



Parallel PISCES Speedup

Distribution of Run Time



Parallel PISCES Run Time as an Accumulation of Parts for 21 X 21 Diode

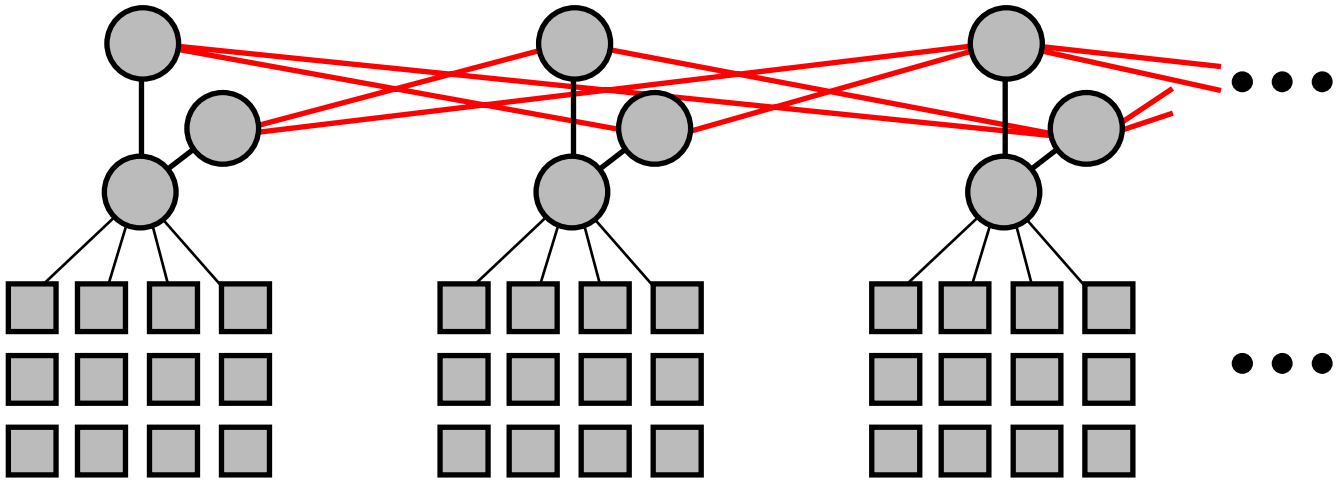
What happened?

- Boundary conditions changed
- PISCES:
  - I spent eighteen months porting PISCES
  - Meanwhile, Pinto and Rafferty kept working
  - Parallel PISCES was obsolete before it was finished
- Computers:
  - Good: iPSC out-performed Sun and VAX
  - Bad: iPSC roughly matched the Convex C-1
- You get what you pay for!
- Bottom Line:
  - Parallel PISCES was only used to generate curves for my thesis.

Application War Story #2

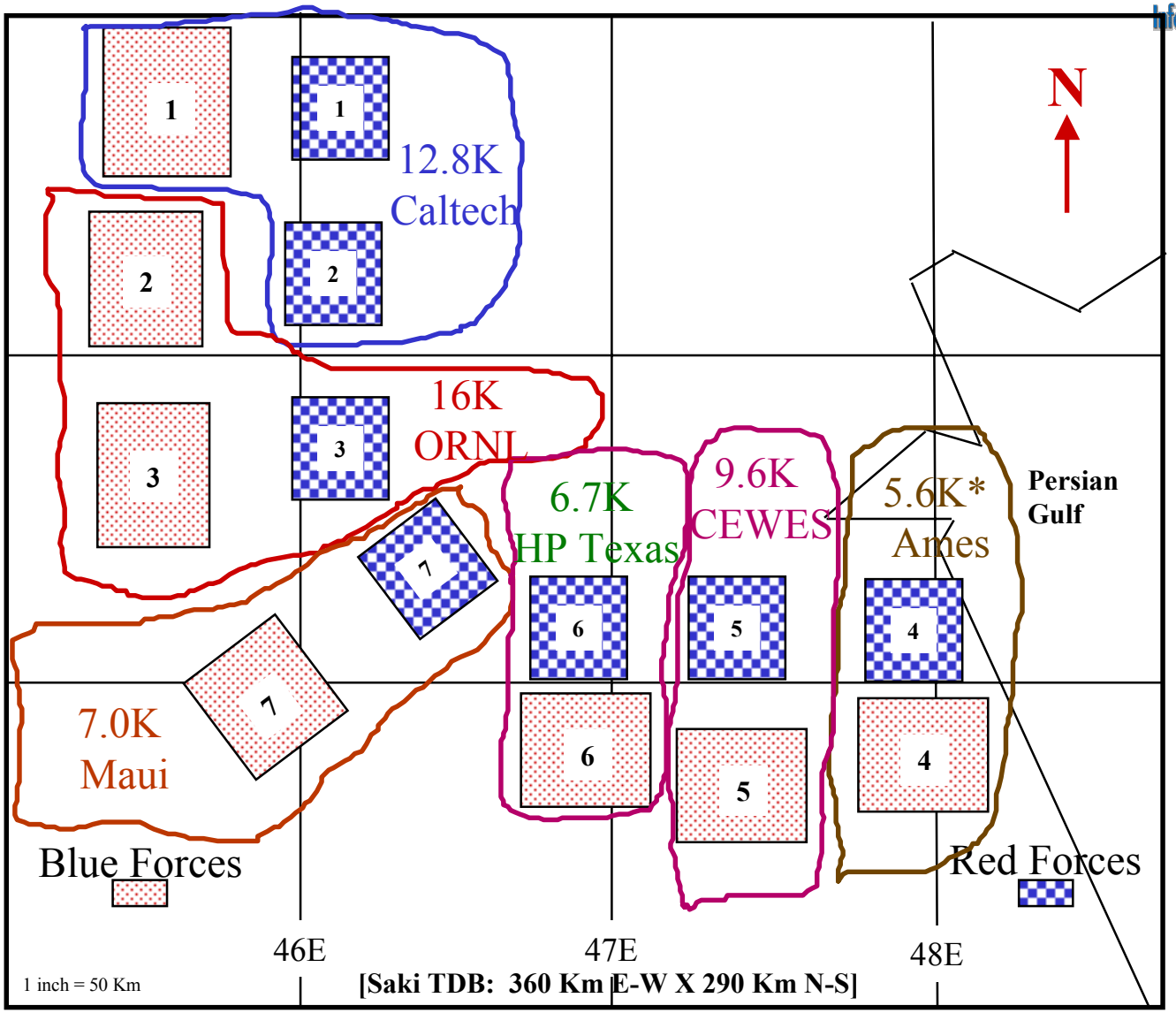
- DARPA project: SFExpress
- Baseline is modSAF
  - Human-in-the-loop simulator for training
  - Bottleneck is communicating state amongst entities
  - Goal is to run 50,000 entities (I.e., tanks)
  - State-of-the-art was ~2000

Full Pathfinder SPP Architecture

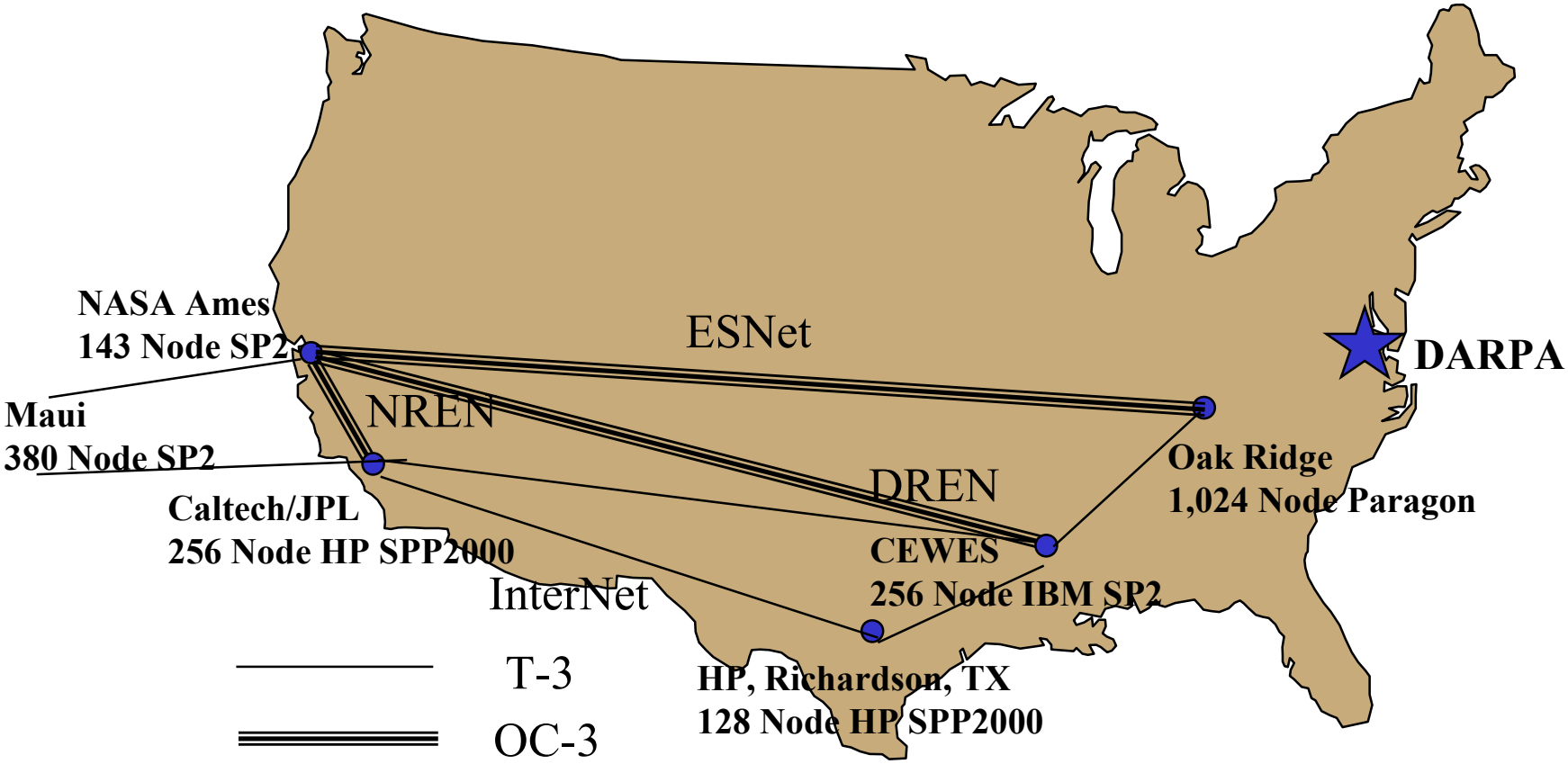


- **Replicated Basic Units To Support Total Entity Count**
- **Independent Communications Within Up-Down Layers**
  - Parallel Operations With SAFSim Services
- **Improve Performance: Use More, Smaller Basic Units**

“50K” Simulation,Scenario V 3.0



Early Grid Application



Demonstrable Scalability

Run Size (Nodes)	81	161	238
Number of Router Triads	3	6	9
Number of SAFSim Nodes	60	120	180
Number of Simulated Vehicles	4,327	8,529	12,915
Primary Busy Fraction	0.188±0.038	0.189±0.018	0.207±0.035
Pop-Up Busy Fraction	0.025±0.015	0.025±0.007	0.027±0.014
Pull-Down Busy Fraction	0.030±0.022	0.026±0.018	0.031±0.016
Primary Receive Time [msec]	0.560±0.115	0.537±0.057	0.587±0.089
SAFSim Comms. Fraction	0.023±0.021	0.024±0.011	0.030±0.037
SAFSim Receive Time [msec]	1.191±2.200	0.978±0.912	1.526±2.652

Table 1: Router and SAFSim performance measures for a sequence of runs of the Maui High Performance Computing Center's IBM SP2.

What happened?

- STOW 97 “diminished down expectations”
  - STOW 97 ran around 5,000 entities
  - SFExpress achieved 100,000 entities
- modSAF development continued independently
- We never changed the mainstream code
- SFExpress had little impact ☹

Is A Pattern Emerging?

- Parallelization efforts succeeded, yet had little impact
- Critical flaw was they were not the mainstream code, and could never catch up
- Lesson Learned!
  - Don’t just do research projects and stunts
  - Work real codes and real problems
  - The Apollo project was a research project!!!

My Programming Odyssey: My Youth

- SAIL – Undergraduate course work
- X86 assembly – “firmware”
- Hexadecimal – career as EE peaked ☺
- Pascal – Masters course work
- Ada – First “research project” language
- f77 – language for PISCES
- f77 + Intel message passing (isend/irecv)

My Programming Odyssey: Post Doctoral Work

- C – good for bit bashing
- CAL – maximum performance
- CMF – TMC's variant of HPF
- AC – Bill Carlson's "vector C" for CM-5
- MPL - MasPar
- AC – Bill Carlson's early version of UPC
- MS Office ☹

My Programming Odyssey: Today

- f77 and MPI
- Why?
  - f77 compilers generate fast code
  - MPI is the lowest common denominator
- Tomorrow?
  - Maybe Java + f77 kernels + MPI
  - Right programming model for a Beowulf

What have I learned About Programming Models?

- Message passing tedious and error prone
- Shared memory is better
  - Simpler programming abstraction
  - Lower latency when supported in H/W
  - One can evolve code like on vector machines
- Both usually lead to the same end-point
  - Exploit locality
  - Minimize inter-processor interaction

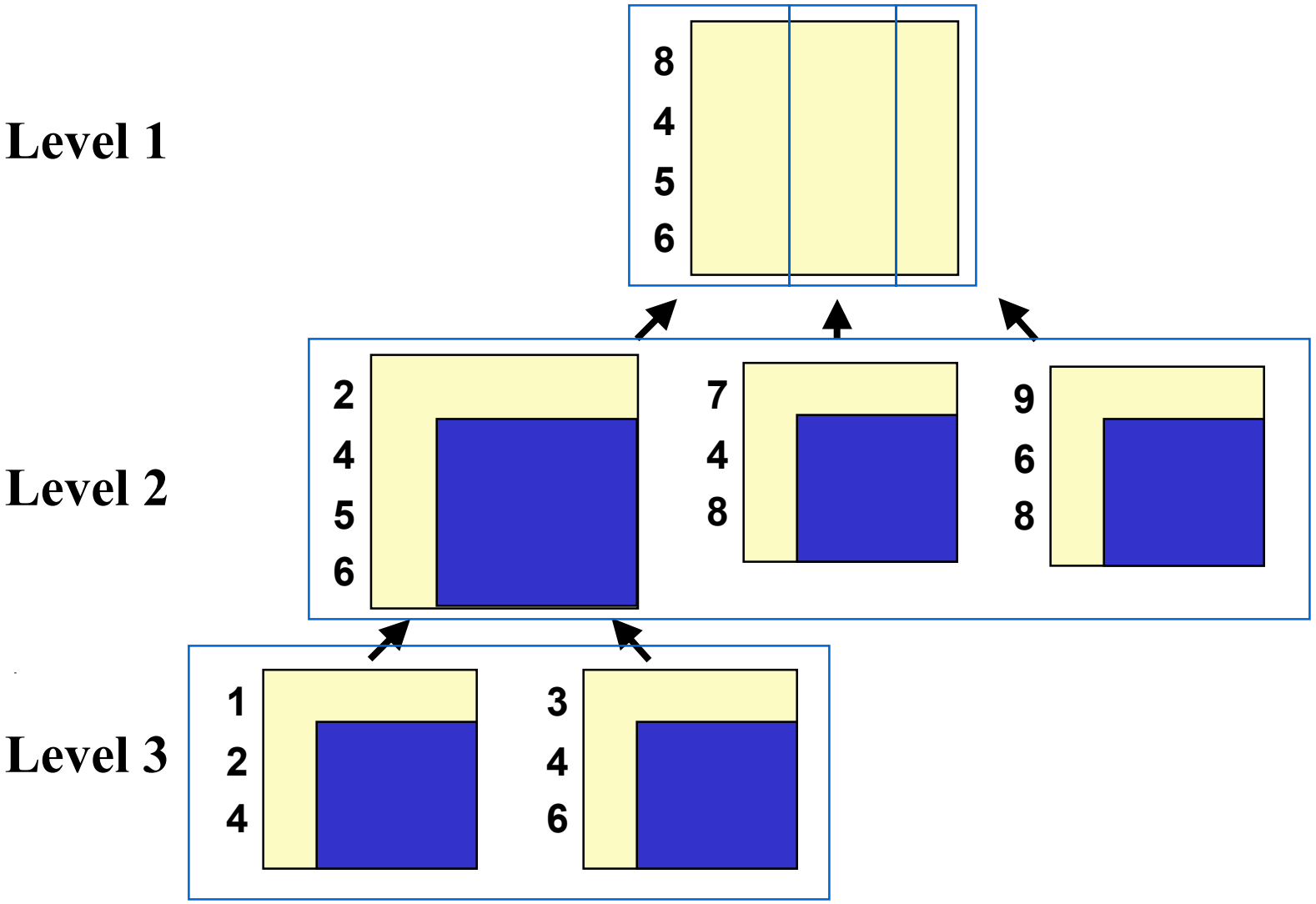
Nevertheless …

- Code usually outlives any one machine
- Machine models keep changing
  - PVP
  - SIMD
  - NUMA
- Therefore, use lowest common denominator
  - F77 and/or C
  - MPI
  - Maybe a C++ or Java shell
  - Perhaps even Python?

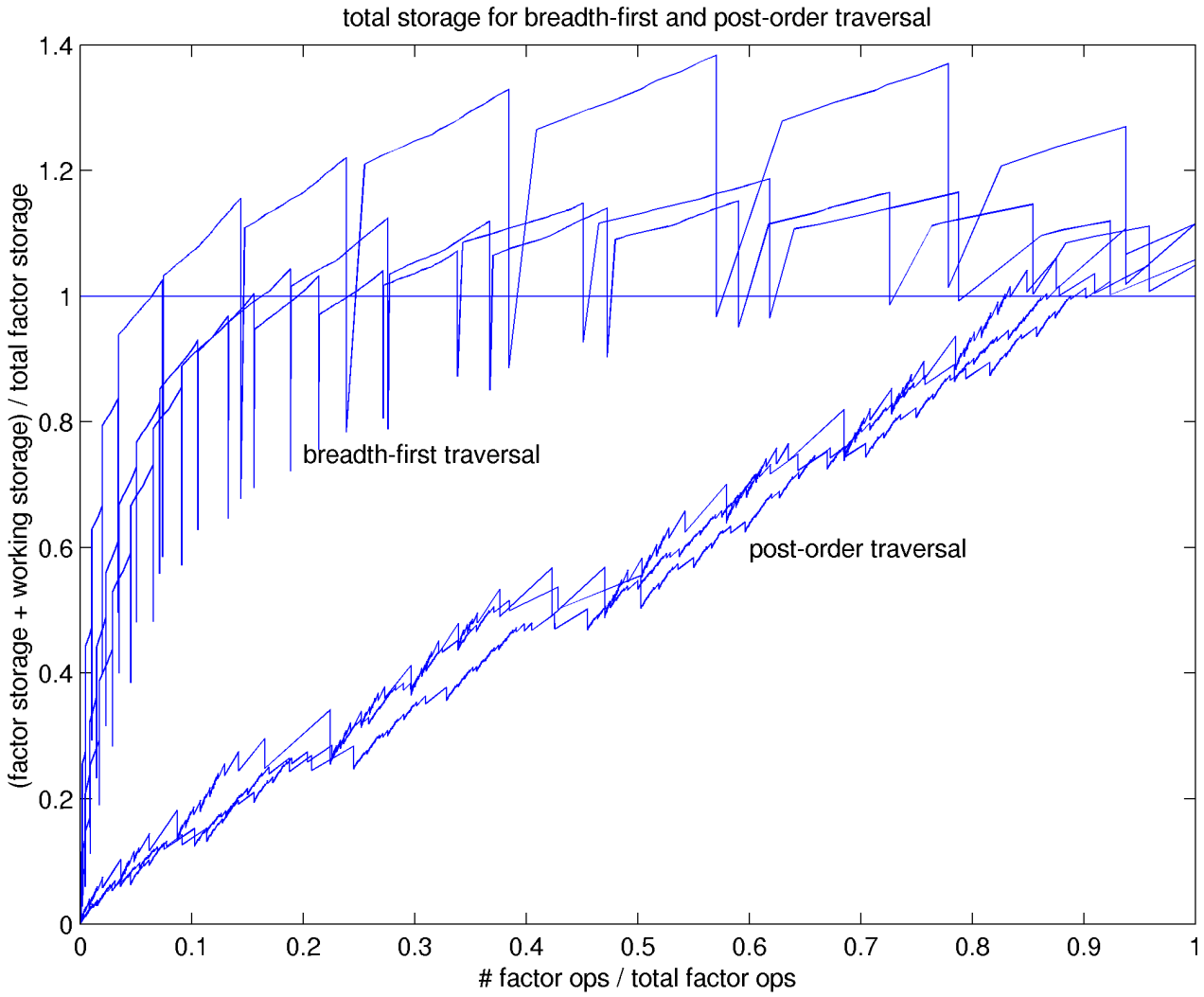
What have I learned About Productivity?

- Productivity is a function of many things
  - Familiarity
  - Abstraction (AC and MPL were not virtual)
  - Correlation between programming model and H/W
- Intellectual Risks Compound
  - New mathematical algorithms
  - Parallel Processing
    - Distribution of work and data
    - Coordination
  - Better addressed one at a time

Performance Frustration: Sparse Matrix Factorization

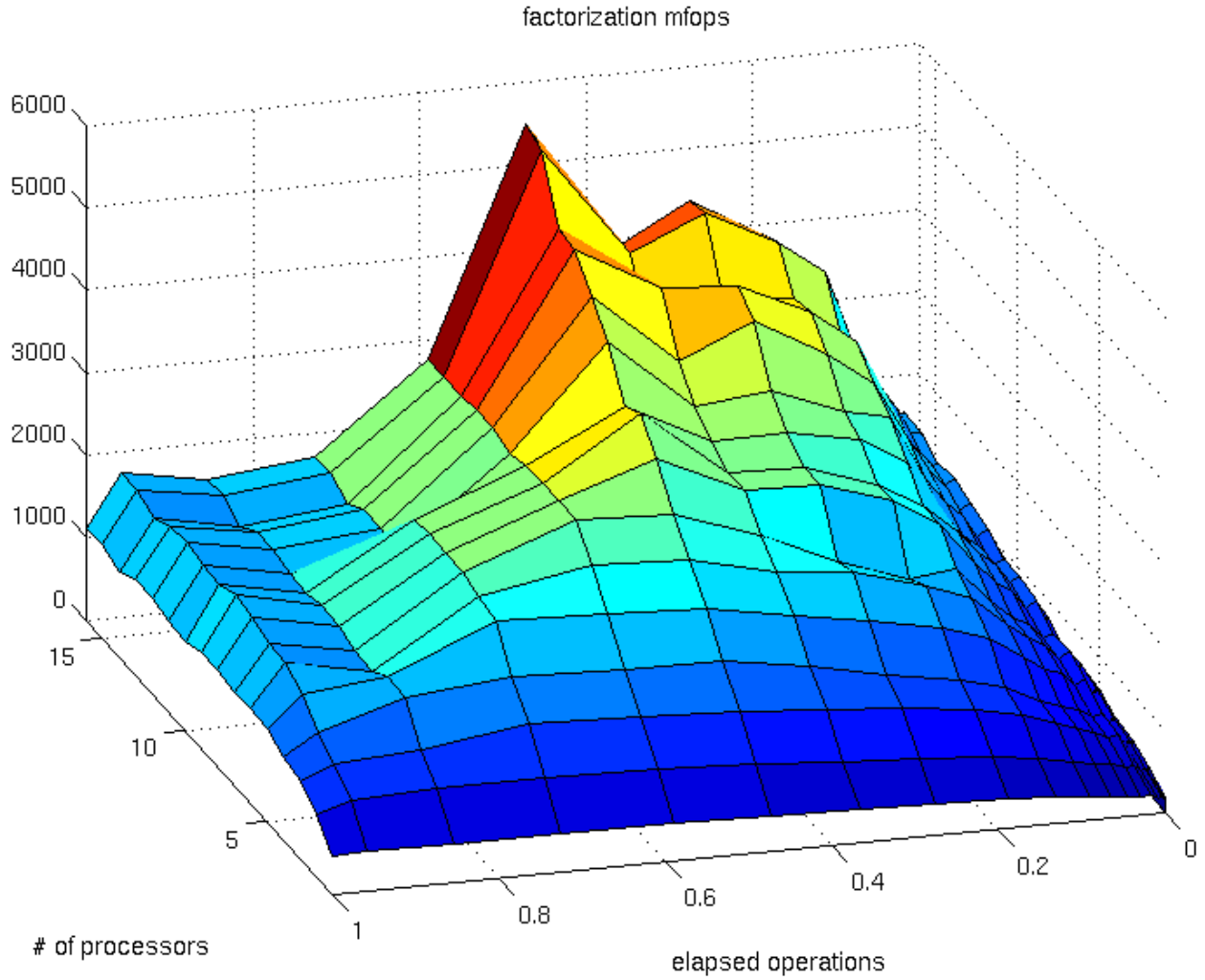


Sparse Matrix Factorization Storage Traces



Hood Performance on O2K

Information Sciences Institute



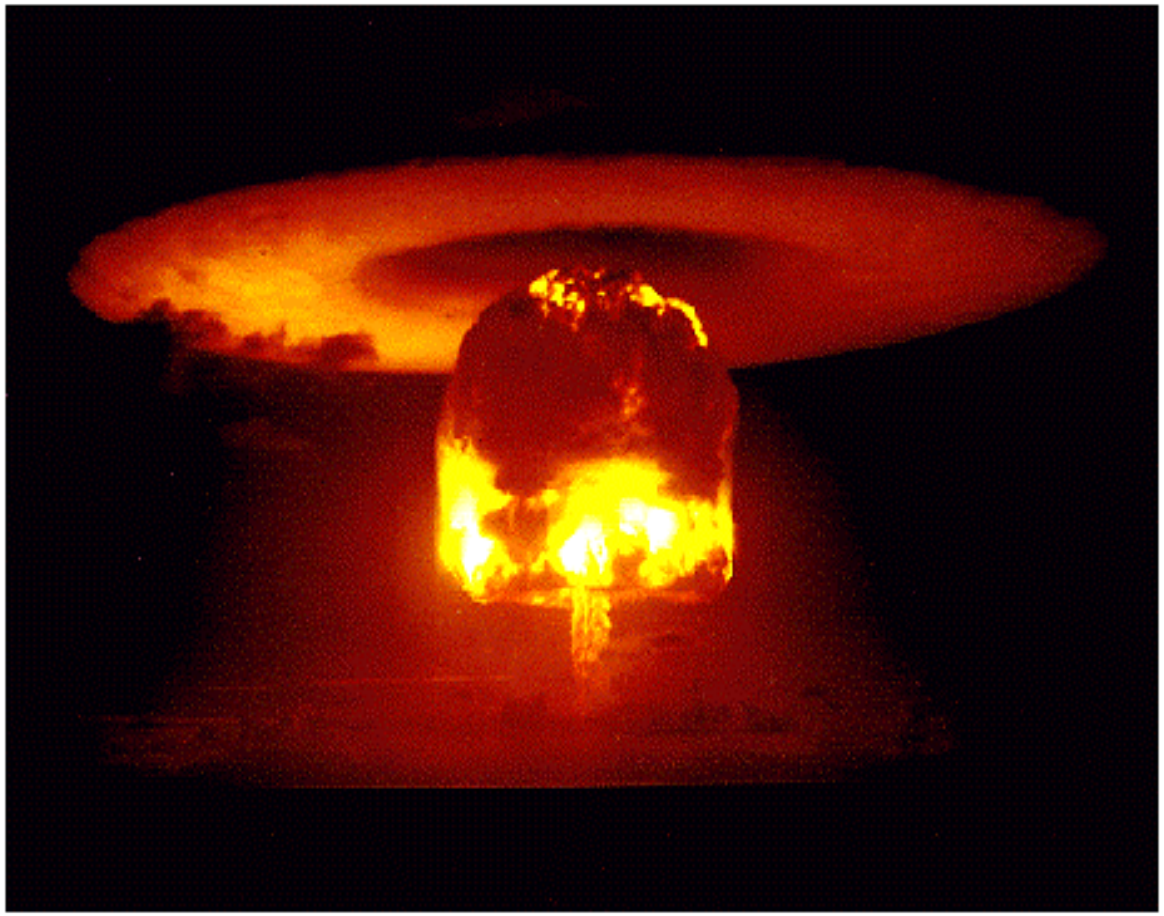
Bottom line on tools?

- printf and etime
  - Lowest common denominator
- Of course I've seen better
  - Totalview
  - FlashPoint
- I'm frustrated!!!

Vision Thing

Information Sciences Institute

Start with a motivating problem!



ASCI Got This Right

Up and to right chart

To Serve Man

- Are they missionaries here to save us?
- Is it a cookbook?
- The spooks can't articulate their problems

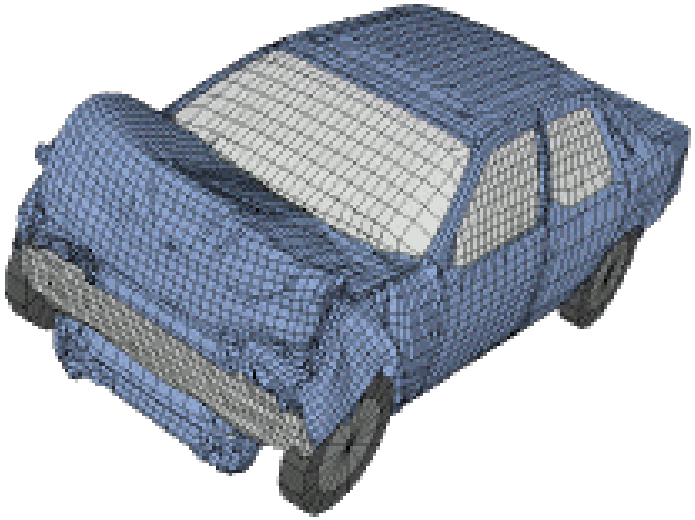
“Malaise”

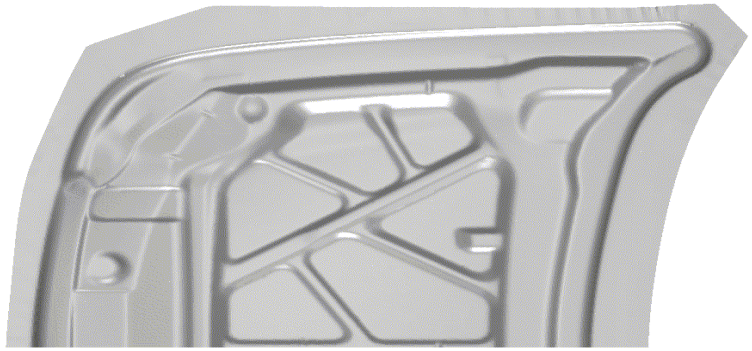
- Parallel processing is hard, if not impossible
- Many people have given up
- Others have no vision or energy

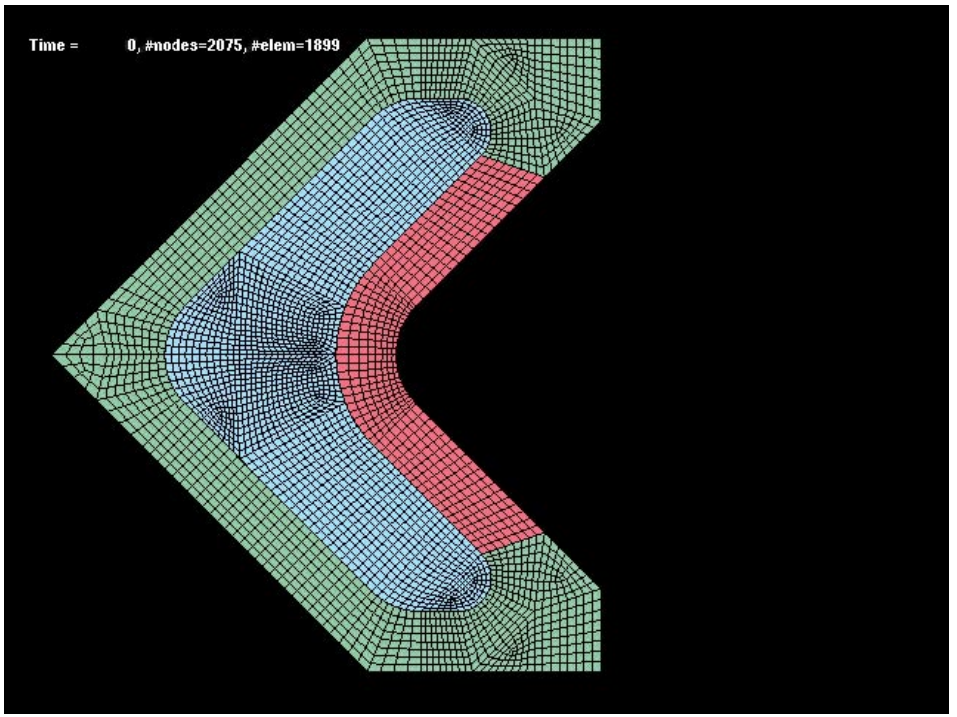
Be “Like Rick”

- You need to be an eternal optimist
- You need to have big dreams
- You need to stick with projects to the end and deliver (E.g. MPI)

Mechanical Dynamics Example





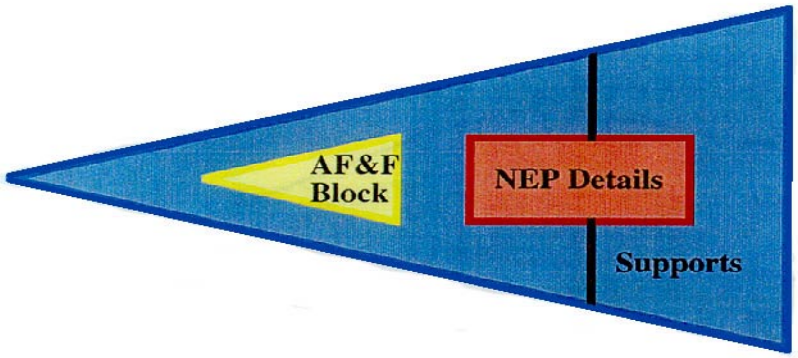


John Hallquist's Vision for How to Exploit Additional Cycles

- First order-of-magnitude of effective performance
  - Realistic models
- Next two orders-of-magnitude
  - Automatic design space exploration
- Next two …
  - Non-ideal material properties
- Next one …
  - Over lunch instead of over night
- Next one …
  - Over a smoke ☺

How's Their Performance?

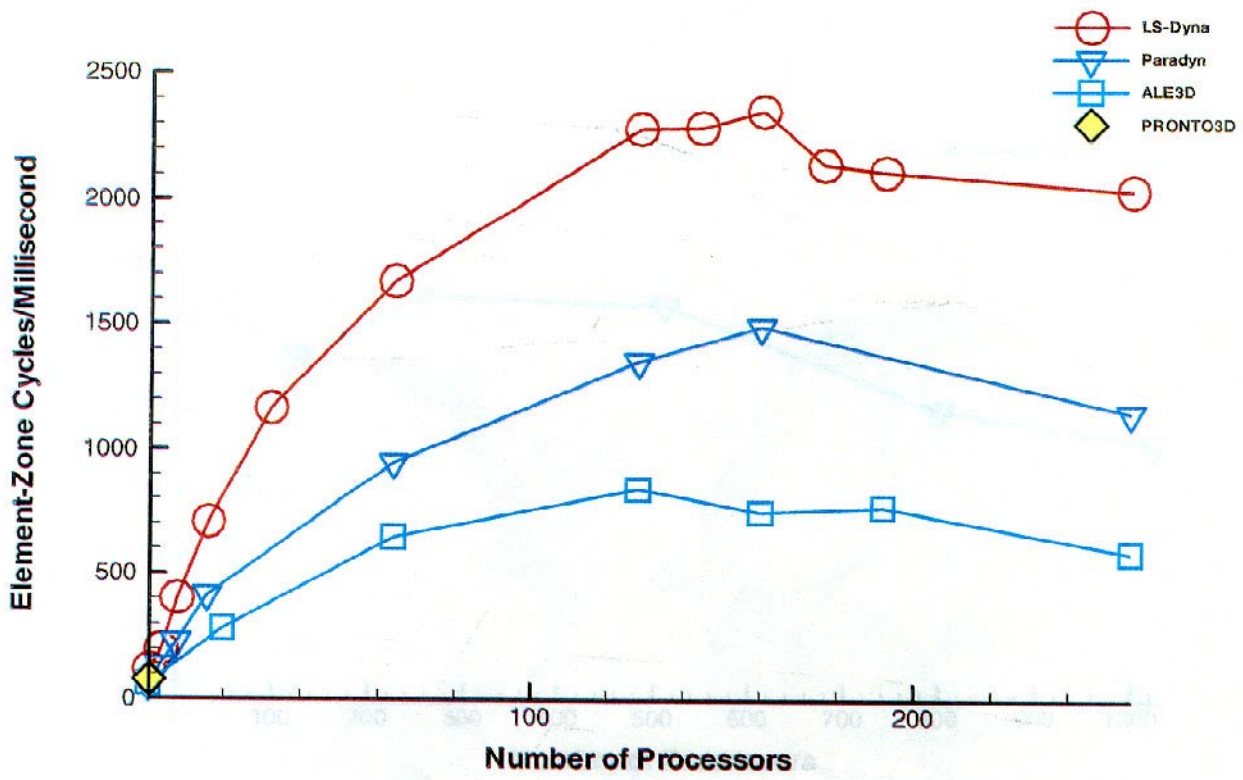
Due diligence from LLNL



62 Materials

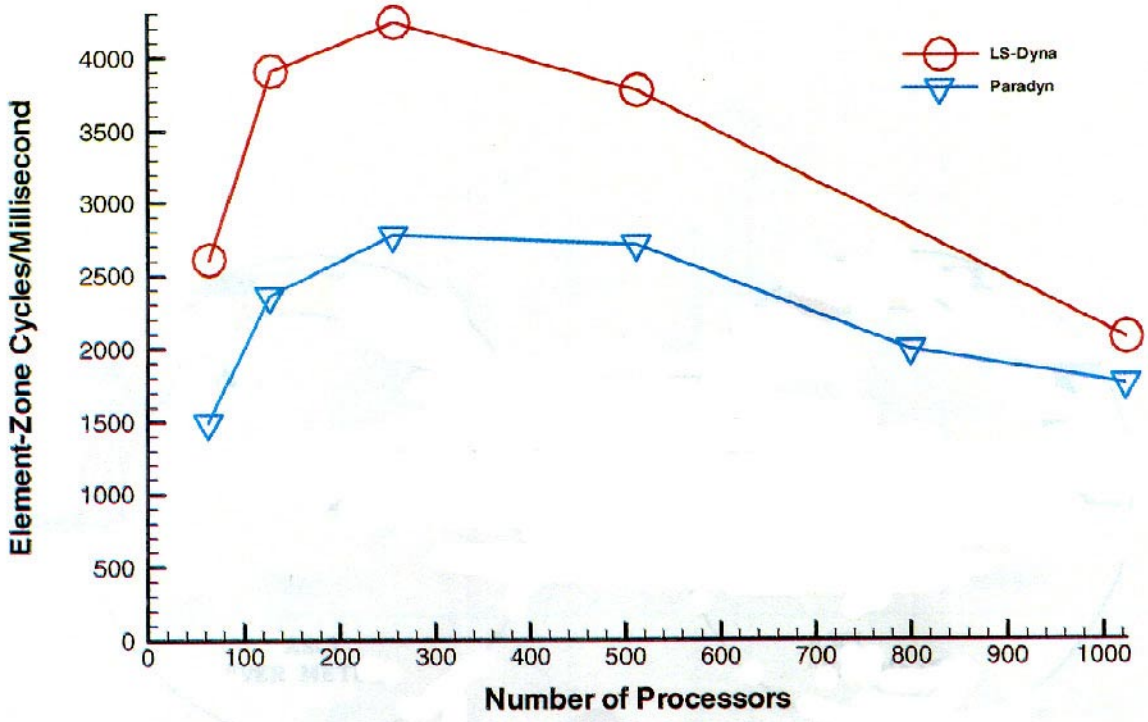
Coarse Model	**Dense Model**
259,990 nodes	1,166,444 nodes
208,293 elements	1,003,922 elements

Fixed Speedup for Small Model

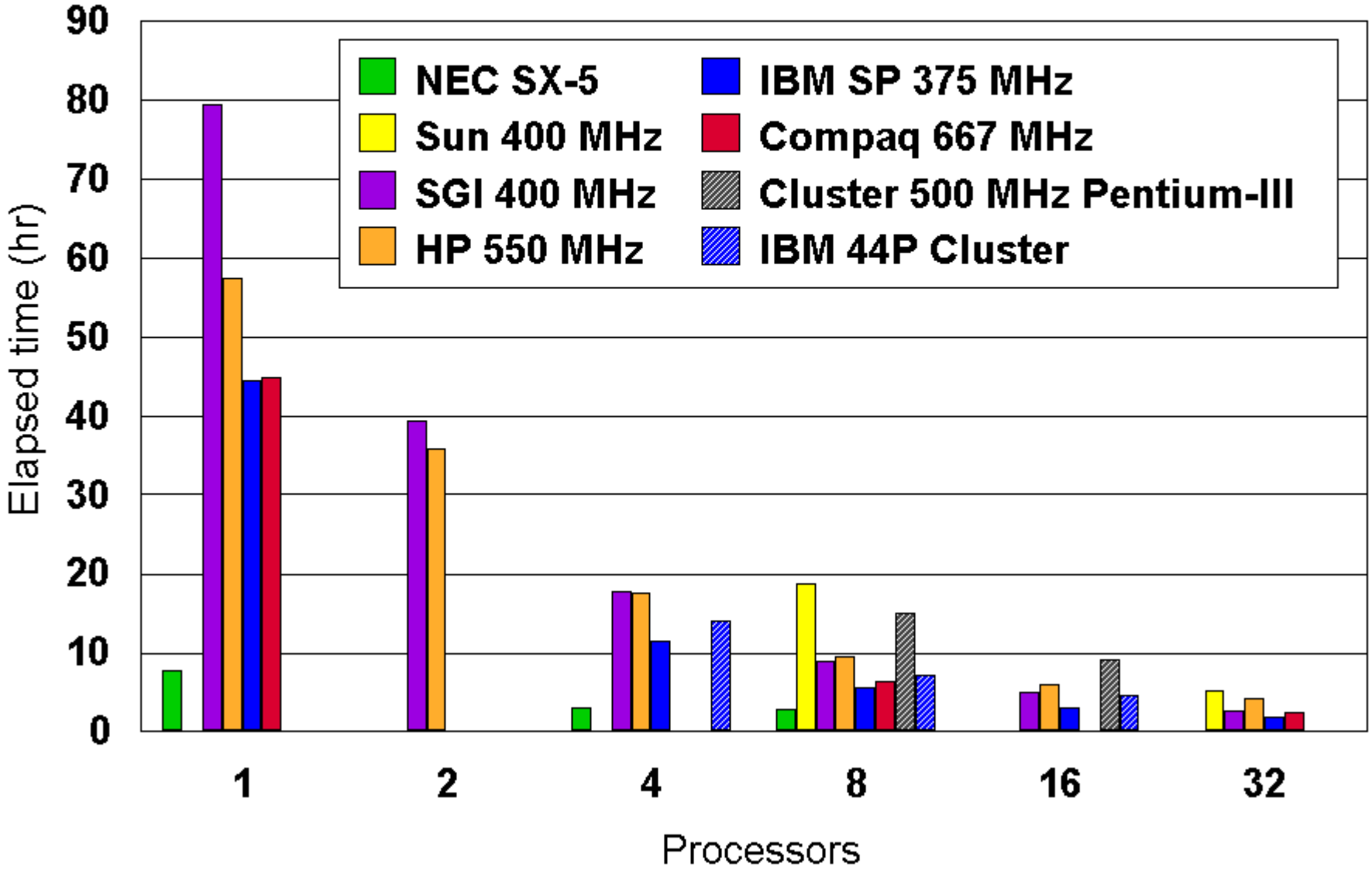


Fixed Speedup for Big Model

Information Sciences Institute



LS-DYNA Performance
80 msec NCAC Neon Model
Data from the 2000 LS-DYNA Conference



Normalized Performance

- Note, y-axis not labeled in Flop/s!!!
- Detour to IBM slide …

Why so slow?

- Irregular grids
  - NP-complete partitioning problem
- Adaptive grids
  - Keep revisiting it
- Contact search
  - Giant parallel sort
- Localized physics
  - Load Imbalance
- Implicit time steps?
  - God help you ☹

Future Performance?

- We're struggling with today's systems
  - Electrical Engineers gave up
  - Mechanical Engineers OK on small systems
  - Material Scientists OK today, but …
- How effective can Blue Gene/L be?
- What would Gene Amdahl say?

How About the ES40?

- The Japanese kicked our collective butts ☹
- It appears they worked backwards from an attractive application
  - think Kyoto
- They did not say “woe is me, I can only have a big PC cluster”
- They maintained focus and $$$
  - Not just a criticism of politicians
  - Research community runs from one fad to the next

What I Told You About Parallel Applications

- Work outside the mainstream often stays there
  - Parallel PISCES
  - SFExpress
- Need to solve real problems
  - Its not enough to build big machines
  - Its not enough to publish research papers

What I Told You About Programming

- People settle on the lowest common denominator
- Productivity is a function of familiarity
- Shared memory better
  - Allows users to evolve

What I Told You About Vision

- Think Big!
  - ASCI got this right
- Think in terms of real problems
  - Don't need whimsical applications to have a vision
  - Be honest about your performance
- Evolve
  - Allow people to get from here to there
  - Create an ES40 for US science ☺

Backups

What Tom Blank Told Me

- Tell ‘em what you’re going to tell ‘em
- Then tell ‘em
- Then tell ‘em what you told ‘em

Relative Processor Performance

Processor	Total Time (sec.)	Major Newton Routines	
		Assemble	Factor
Intel 310/142	1165	36%	49%
SUN 3/50	554	42%	49%
Convex C-1	103	46%	25%

Time (sec.) and Percentage of Total Time
That PISCES Spent Executing Key Routines
in the Sample Diode Problem.